An Axiomatically Derived Measure for the Evaluation of Classification Algorithms

Fabrizio Sebastiani*
Qatar Computing Research Institute
Hamad bin Khalifa University
PO Box 5825, Doha, Qatar
fsebastiani@qf.org.qa

ABSTRACT

We address the general problem of finding suitable evaluation measures for classification systems. To this end, we adopt an axiomatic approach, i.e., we discuss a number of properties ("axioms") that an evaluation measure for classification should arguably satisfy. We start our analysis by addressing binary classification. We show that F_1, nowadays considered a standard measure for the evaluation of binary classification systems, does not comply with a number of them, and should thus be considered unsatisfactory. We go on to discuss an alternative, simple evaluation measure for binary classification, that we call K, and show that it instead satisfies all the previously proposed axioms. We thus argue that researchers and practitioners should replace F_1 with K in their everyday binary classification practice. We carry on our analysis by showing that K can be smoothly extended to deal with single-label multi-class classification, cost-sensitive classification, and ordinal classification.

1. INTRODUCTION

Classification is an enabling technology of capital importance in nowadays' data science, and plays a central role in countless tasks of practical importance, including text classification, spam filtering, word sense disambiguation, Web search, data mining and knowledge discovery, and others. As in all data-related endeavours, experimental evaluation plays a central role in classification, and the mathematical measure that we adopt is the cornerstone of this evaluation. In the last 20 years the F_1 measure (the harmonic mean of precision and recall – sometimes colloquially termed the "F-score" or the "F-measure") has progressively replaced "accuracy" (the fraction of classification decisions that are correct, which corresponds to the complement of "Hamming distance" or "0-1 loss") as the standard evaluation measure of binary classification in information retrieval (IR), machine learning, data mining, and NLP.

*Fabrizio Sebastiani is currently on leave from Consiglio Nazionale delle Ricerche, Italy.

ICTIR '15, September 27-30, 2015, Northampton, MA, USA
© 2015 ACM. ISBN 978-1-4503-3833-2/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808194.2809449.

In this paper we challenge F_1 and its suitability for evaluating binary classification. To this end we adopt an *axiomatic* approach, i.e., one based on arguing in favour of a number of properties ("axioms") that an evaluation measure for classification should intuitively satisfy. The benefit of this axiomatic approach (which has a rich history in IR – see Section 7) is that it shifts the discussion from the evaluation measures to the axioms, which is like shifting the discussion from a complex combination to its building blocks: once the scientific community has agreed on a set of axioms (the building blocks), it then follows whether a given measure (the combination) is satisfactory or not. After discussing these axioms, we study F_1 and a few other existing measures for binary classification, and find them to be unsatisfactory, in the sense that they all fail to satisfy some of the properties we have argued for. We carry on to propose K, a new evaluation measure for binary classification, which actually consists of a variant of measures ("balanced accuracy", "Youden's index") which have surfaced in the past in classification or related endeavours; we formally prove that K satisfies all the properties we have previously argued for.

Since K can deal with binary classification it can also deal with *multi-label multi-class classification* (MLMCC), i.e., the case in which zero, one, or several from a set $\mathcal{C}$ of available classes (with $|\mathcal{C}| > 1$) can be attributed to a given item. We then go on to show that K can smoothly be extended to deal with *single-label classification* (SLC – i.e., when exactly one class must be chosen from set $\mathcal{C}$, with $|\mathcal{C}| > 1$), with *cost-sensitive classification* (CSC – i.e., when different types of misclassification may have different costs), and with *ordinal classification* (OC – i.e., when such costs are constrained by a linear order defined on $\mathcal{C}$). This shows that K can be used as a unifying measure for all types of classification (binary, MLMC, SL, cost-sensitive, ordinal).

Note that in this paper we only deal with the problem of evaluating "hard" classification (i.e., the crisp assignment of classes to items), and not with evaluating systems that rank items according to their degree / probability of membership in a class ("soft" classification)[1]. For the same reason we disregard (a) measures (such as "precision as a function of recall") that do not depend on the choice of a classification threshold, and (b) measures (such as "precision-recall breakeven point" – see also Section 4) in which the threshold is chosen not by the system but (cryptically enough) by the evaluation software.

[1]Hard and soft classification are sometimes referred to as "autonomous" and "interactive" classification, respectively; see e.g., [15].

The rest of the paper is organised as follows. Section 2 discusses some known measures for evaluating binary classification. In Section 3 we argue in favour of a series of properties ("axioms") that we claim binary classification measures should satisfy, while in Section 4 we show that F_1 and some existing measures for binary classification do not satisfy some of them. Section 5 is devoted to discussing the K measure, and to showing that it does satisfy all of the axioms proposed in Section 3. Section 6 deals instead with extending K to the SLC case, to the cost-sensitive case, and to the ordinal case. Section 7 discusses related work, while Section 8 concludes.

2. KNOWN MEASURES OF CLASSIFICATION EFFECTIVENESS

2.1 Preliminaries

In Sections 2 to 5 we restrict our discussion to binary classification. Let $\mathcal{D}$ be a domain of items, let c be a class, and let $\mathcal{Y}_c : \mathcal{D} \rightarrow \{-1, +1\}$ be the *target function* for c, where -1 and $+1$ indicate non-membership and membership in c, respectively. We denote by $D \subseteq \mathcal{D}$ a nonempty set of items on which the effectiveness of classifiers needs to be evaluated. A pair $\langle D, \mathcal{Y}_c\rangle$ will be called a *test set* for c. We denote by $h_c : \mathcal{D} \rightarrow \{-1, +1\}$ a *classifier* (or *hypothesis*, or *predictor*) for c. We will thus call $\mathcal{Y}_c(d)$ and $h_c(d)$ the *actual label* and the *predicted label* of d for c, respectively.

We will also denote

- by $\overline{c}$ the complement of class c;
- by $\overline{\mathcal{Y}}_{\overline{c}}$ the complement of target function $\mathcal{Y}_c$, defined as the target function for $\overline{c}$ such that $\mathcal{Y}_c(d) = -\overline{\mathcal{Y}}_{\overline{c}}(d)$;
- by $\overline{h}_{\overline{c}}$ the complement of classifier h_c, defined as the classifier for $\overline{c}$ such that $h_c(d) = -\overline{h}_{\overline{c}}(d)$.

Note that $\mathcal{Y}_c$ and $\overline{\mathcal{Y}}_{\overline{c}}$ are essentially the same function, although the former is framed in terms of c and the latter is framed in terms of $\overline{c}$. For instance, if c stands for ProRepublican and $\overline{c}$ stands for ProDemocrat, the same items that belong to ProRepublican according to $\mathcal{Y}$ also belong to ProRepublican according to $\overline{\mathcal{Y}}$, but according to $\mathcal{Y}$ they are positive examples (of ProRepublican, which is the "positive class" for $\mathcal{Y}$) while according to $\overline{\mathcal{Y}}$ they are negative examples (of ProDemocrat, which is the "positive class" for $\overline{\mathcal{Y}}$). The same we have said of target functions $\mathcal{Y}$ and $\overline{\mathcal{Y}}$ applies to classifiers h_c and $\overline{h}_{\overline{c}}$ too.

Some special classifiers that we will refer to are

- the *trivial acceptor* h_c^{acc} (i.e., the classifier that attributes class c to every item);
- the *trivial rejector* h_c^{rej} (i.e., the classifier that attributes class $\overline{c}$ to every item);
- the *perfect classifier* h_c^{perf} (i.e., the classifier that attributes the correct label to every item);
- the *pervert classifier* h_c^{perv} (i.e., the classifier that attributes the wrong label to every item);
- the *random classifier* h_c^{rand} (i.e., the classifier which only takes random classification decisions)[2].

[2]"The random classifier" is actually an abstraction, since there is no unique such classifier; when speaking of h_c^{rand} we will thus be interested in the "average behaviour" of all possible classifiers, i.e., in the expected value of h_c^{rand}.

By TP, FP, FN, TN, we denote the numbers of true positives, false positives, false negatives, and true negatives for class c, as determined by the triple $\langle D, \mathcal{Y}_c, h_c\rangle$. By $AP = TP + FN$ and $AN = TN + FP$ we denote the number of *actual positives* and *actual negatives*, while by $PP = TP+FP$ and $PN = TN+FN$ we denote the number of *predicted positives* and *predicted negatives*, respectively.

We will denote by $M(D, \mathcal{Y}_c, h_c)$ a measure for evaluating the effectiveness of a classifier h_c as applied to a dataset D labelled according to target function $\mathcal{Y}_c$. Note that $M(D, \mathcal{Y}_c, h_c)$ is essentially a function of the two variables TP and TN, since AP and AN are constants for a given pair $\langle D, \mathcal{Y}_c\rangle$, i.e., they are not under the control of the experimenter, and since $FP = (AN - TN)$ and $FN = (AP - TP)$. In this paper we will take M to be a measure of accuracy, and not of inaccuracy, so we will always assume that higher values are better.

One assumption we will make in the rest of this work is that, for each test set $\langle D, \mathcal{Y}_c\rangle$, a *cost vector* $\Lambda(D, c) = (\lambda_1, \ldots, \lambda_{|D|})$ is known in advance, where $\lambda_i > 0$ denotes the cost of misclassifying item d_i for class c, and where $\sum_{i=1}^{|D|} \lambda_i = |D|$. This assumption is not restrictive. For instance, we might want to impose that $\lambda_i = 1$ for all $d_i \in D$, which covers the most frequent case in which all items have the same importance, an assumption which underlies common evaluation measures such as accuracy, F_1, and many others; but other choices are possible, in which different documents are deemed of different importance[3]. Note that, as we have specified, λ_i must be strictly higher than zero for all items d_i; this formalizes the intuition that, when it comes to evaluation, "no item is worthless". In the rest of this paper we will only address measures of the first type, i.e., characterized by the "$\lambda_i = 1$ for all $d_i \in D$" assumption; anything we say can be straightforwardly extended to the case in which this assumption is relaxed.

2.2 Measures for evaluating classification

Table 1 lists a number of "simple" evaluation measures for binary classification that have been proposed or talked about over the years, while Table 2 lists a number of "combined" such measures.

"Simple" measures (called "partial measures" in [3]) involve only two (adjacent) cells of the contingency table and only one between FP and FN. "Combined" measures involve more than two cells of the contingency table and both FP and FN, and often result from the combination of two simple measures, one involving FP and the other involving FN. All of them are expressed as ratios, where the denominator is a certain population of items and the numerator is the part of that population that bears some significance to the behaviour of the system. Since any plausible evaluation measure must come to terms with the ability of the system to avoid *both* false positives and false negatives, simple measures are usually not employed on their own but only as building blocks of combined measures. Note that the same measure may have several alternative names, due to the fact that it may have independently originated in several fields

[3]In order to implement *cost-sensitive classification* [10] we might want to impose, e.g., that $\lambda_i = k_p$ for all $d_i \in AP$ and $\lambda_i = k_n$ for all $d_i \in AN$, where k_p and k_n are two different constants (normalized in such a way that $\sum_{i=1}^{|D|} \lambda_i = |D|$). However, in Section 6.2 we will see a different method of dealing with CSC, which does not require setting different values of λ_i for the items in AP and AN.

Symbol	Name	Formula	Note
ρ	Recall (aka "Sensitivity", "True Positive Rate", "Hit Rate")	$\frac{TP}{TP+FN}$	(1-FNR)
FNR	False Negative Rate (aka "Miss Rate")	$\frac{FN}{TP+FN}$	(1-ρ)
π	Precision (aka "Positive Predicted Value")	$\frac{TP}{TP+FP}$	(1-FDR)
FDR	False Discovery Rate (aka "False Alarm Rate")	$\frac{FP}{TP+FP}$	(1-π)
ϕ	Fallout (aka "False Positive Rate")	$\frac{FP}{FP+TN}$	(1-σ)
σ	Specificity (aka "True Negative Rate", "Inverse Recall")	$\frac{TN}{FP+TN}$	(1-$\mathcal{Y}_c$)
NPR	Negative Predicted Value (aka "Inverse Precision")	$\frac{TN}{FN+TN}$	$(1-\epsilon)$
ϵ	Elusion [20, p. 55]	$\frac{FN}{FN+TN}$	(1-NPR)

Table 1: "Simple" measures computed from a contingency table (alternative names common in disciplines other than IR are also given).

Symbol	Formula
Acc	$\frac{TP+TN}{\lvert D\rvert}$
DOR	$\frac{TP\ TN}{FN\ FP}$
$LAM\%$	$\frac{1}{\log \frac{\frac{1}{2}\log \frac{FP\ FN}{TP\ TN}}{1-\frac{1}{2}\log\frac{FP\ FN}{TP\ TN}}}$
ASP	$\frac{TP^2}{AP\ PP}$
MCC	$\frac{(TP\ TN)-(FP\ FN)}{(AP\ AN\ (TP+TN)\ (FP+FN))^{\frac{1}{2}}}$
F_1	$\frac{2TP}{2TP+FP+FN}$

Table 2: Common "combined" measures computed from a contingency table.

far from each other (e.g., IR, signal detection, diagnostic testing).

2.2.1 *"Simple" measures*

Among the simple measures listed in Table 1, precision (π), recall (ρ), fallout (ϕ), and specificity (σ) are historically the most important. Recall has been the universally adopted way to measure the ability of the system to avoid false negatives. Instead, the ability of the system to avoid false positives has been measured in various ways (precision, fallout, specificity); in IR fallout was the measure of choice in the '60s, but was gradually replaced by precision, while other fields such as e.g., epidemiology, have instead always relied on specificity, the complement of fallout. Note that, while fallout and specificity are independent of recall (since they use non-overlapping parts of the contingency table), precision is not.

2.2.2 *"Complex" measures*

Accuracy (Acc), the fraction of classification decisions that are correct, has been for many years the measure of choice in machine learning and statistics, mostly because of its simplicity. Accuracy is little used in text classification and other endeavours characterised by high imbalance (typically meaning that $AP \ll AN$), since in this case the trivial rejector trivially obtains high values; F_1 is usually the measure of choice in these cases. As a measure of binary classification performance in diagnostic testing, Glas et al. [13] proposed *diagnostic odds ratio* (DOR), defined as $DOR = (TP \cdot TN)/(FP \cdot FN)$; the same measure is then used in [7] for measuring spam filtering performance. Actually, one binary classification measure popular in the spam filtering community is the *logistic average misclassification percentage* ($LAM\%$ – see e.g., [8]); differently from other measures discussed in this paper, $LAM\%$ is a measure of ineffectiveness, and not one of effectiveness, i.e., low $LAM\%$ values are better. Other measures that have been put forward in the past are *average set precision* (ASP; see [14]), originally proposed in the context of the TREC filtering track; the *Matthews Correlation Coefficient* (MCC; see [18]), which originated within biochemistry; and, of course, F_1.

3. AXIOMS FOR CLASSIFICATION

We approach the issue of how to evaluate binary classification in an *axiomatic* way, i.e., by (a) arguing for a number of properties that an evaluation measure for binary classification should satisfy, (b) studying existing evaluation measures in terms of whether they satisfy these properties or not, and possibly (c) synthesizing new measures that do satisfy them. An advantage of this method is that research on evaluation measures may proceed, rather than by challenging previously proposed measures, by challenging previously proposed *axioms*, and by possibly arguing in favour of new ones. Once the scientific community has converged on a given set of axioms thanks to this process, the suitability of existing measures is immediate to ascertain, and the synthesis of measures that satisfy this set is made easier.

3.1 The axioms

We argue that a function $M(D, \mathcal{Y}_c, h_c)$ that measures the effectiveness of binary classifiers h_c should obey the following axioms. We will often write $M(\mathcal{Y}_c, h_c)$ instead of $M(D, \mathcal{Y}_c, h_c)$ when the first argument is clear from the context.

AXIOM 1. ***Strict Monotonicity (MON)****. For any test set $\langle D, \mathcal{Y}_c\rangle$, and for all classifiers h_c and h'_c such that h'_c differs from h_c only for the label attributed to a single item $d \in D$, wrong for h_c and correct for h'_c, it holds that $M(\mathcal{Y}_c, h_c) < M(\mathcal{Y}_c, h'_c)$.* □

MON enforces the notion that in no case the evaluation measure can be indifferent to the fact that a given classification decision is correct or wrong; that is, the monotonicity of M should be *strict*. **MON** is a direct consequence of the assumption (see Section 2.1) that for no item d_i the cost of misclassifying d_i can be zero.

Note that what **MON** says in practice is that, given D and $\mathcal{Y}_c$, the measure should be sensitive to *both* the number FP of false positives and the number FN of false negatives. It does *not* state that it should be sensitive to the values of precision and recall; these latter are *derived* notions (i.e., functions of the contingency table), while FP and FN are *primitive* elements of the same table.

In [3] **MON** is called the "growing quality constraint". A consequence of **MON** is that M can achieve its maximum value *only* when the predicted label equals the actual label for all $d_i \in D$ (i.e., when h_c is the "perfect classifier"). [3] calls this the "best system constraint"; we do not list this as a separate axiom since it is a direct consequence of **MON**, and since we deem **MON** mandatory anyway. Analogously, a consequence of **MON** is that M can achieve its minimum value *only* when the predicted label is different from the actual label for all items in the set (i.e., when h_c is the "pervert classifier").

Axiom 2. ***Continuous differentiability (CON)**. For any test set $\langle D, \mathcal{Y}_c\rangle$, M is a continuously differentiable function of TP and TN.* □

We have argued in Section 2.1 that M is essentially a function of TP and TN; **CON** states that it should be a member of the class C^1 of *continuously differentiable* functions. That is, both M and its first derivative should be continuous in both TP and TN. To see the rationale of this, imagine that TP and TN were "masses" instead of "counts"; this requirement has the goal of ensuring that M should behave "reasonably", i.e., respond minimally (i.e., smoothly) to minimal variations of TP and TN, throughout the domain on which it is defined. The intuition behind this axiom is that we want small variations in the contingency table to bring about variations in the value of M that are small themselves.

Axiom 3. ***Strong Definiteness (SDE)**. M is defined for any test set $\langle D, \mathcal{Y}_c\rangle$ and for any classifier h_c.* □

The rationale of **SDE** is fairly obvious, i.e., we want our evaluation measure to always return an answer insofar as the situation being evaluated (the test set, the classifier) is a legitimate one.

Axiom 4. ***Weak Definiteness (WDE)**. For any test set $\langle D, \mathcal{Y}_c\rangle$, and for any classifiers h_c and h'_c, M is defined for h_c iff it is defined for h'_c.* □

This is a weaker definiteness requirement than **SDE**, which acknowledges the fact that sometimes M might not be defined (e.g., because the measure is defined as a ratio and the denominator is zero). The rationale of **WDE** is that, when and if the evaluation function is not defined, it must be such because of the problem itself, and not because of the classifier we want to evaluate. That is, if the function is defined for one classifier, it must be defined for all classifiers, since we cannot afford to comparatively evaluate classifiers defined on the same class c and find out that they are incomparable. This is well explained by Robertson in the context of binary retrieval [22]:

> (...) Such difficulties are almost bound to occur if ratios are used, and there is no hope of comparing results if ratios are not used. (...) But I would like to distinguish between the two cases. The case [in which there are no actual positives] refers to a particular type of question, and does not depend on the test results. If such questions are used to test systems, they can be treated separately from the rest. But the case [in which there are no predicted positives] might occur in answer to any question; to leave such cases out of the averages would be to distort the results.

Axiom 5. ***Fixed Range (FIX)**. The set of values $[\alpha, \beta]$ on which M ranges is fixed, and independent of the test set $\langle D, \mathcal{Y}_c\rangle$.* □

The rationale of **FIX** is that, in order to be able to intuitively judge whether a given value of M means high accuracy or low accuracy, we need to known what values M ranges on, and these values must be independent of the problem setting. That this range is constant regardless of item set D and class c is a necessary condition for us to be able to *immediately* interpret the meaning of a given value of M. (it is not a sufficient condition, though; more on this in the next paragraphs)

Axiom 6. ***Robustness to Chance (CHA)**. It holds that $E[M(\mathcal{Y}_c, h_c^{rand})] = \gamma$, where $E[\cdot]$ indicates "expected value" and γ is a constant independent of the test set $\langle D, \mathcal{Y}_c\rangle$.* □

CHA says that the expected value of M for the random classifier should always be the same, irrespective of class frequency and other factors. Its rationale is allowing the experimenter to fully appreciate a result by correctly placing it into the context of what the random classifier is expected to return. In other words, once that for a given classifier h_c we are told that $M(\mathcal{Y}_c, h_c) = a$, we should be in a position to know how much of a is due to the chance agreement between test set and prediction, and how much is instead due to the true insight of h_c. **CHA** says that a good measure should allow to easily factor out, or discount, the chance effect from one's results.

Axiom 7. ***Robustness to Imbalance (IMB)**. It holds that $M(\mathcal{Y}_c, h_c^{acc}) = k_1$ and $M(\mathcal{Y}_c, h_c^{rej}) = k_2$ for any test set $\langle D, \mathcal{Y}_c\rangle$ such that $AP > 0$ and $AN > 0$, where k_1 and k_2 are two constants independent of $\langle D, \mathcal{Y}_c\rangle$.* □

The rationale of **IMB** is similar to that of **CHA**: trivial classifiers should obtain the same fixed values k_1 and k_2 for all test sets, so that the effectiveness $M(\mathcal{Y}_c, h_c)$ of a given classifier is actually determined by where it falls in the $[\max(k_1, k_2),\ \beta]$ interval, rather than in the $[\alpha, \beta]$ interval discussed for **FIX**. If k_1 and k_2 are the same for all grounds truths, the results returned by M are immediately interpretable, and the experimenter may more easily appreciate the real effectiveness of a classifier.

That k_1 and k_2 are the same for all grounds truths means in particular that they are the same for every level of imbalance. Therefore, a measure that satisfies **IMB** can be meaningfully used *for balanced and imbalanced datasets alike*, and the experimenter can use it without worrying what the level of imbalance in the test set is. This is a striking contrast to the current situation, in which accuracy tends to be considered *the* measure for balanced sets, while F_1 tends to be considered *the* measure for imbalanced sets, a dichotomy that seems unscientific, since it dodges the question as where the threshold between balance and imbalance lies.

The case in which $AN = 0$ is obviously excluded from consideration in this axiom, since in this case the trivial acceptor h_c^{acc} is indeed the perfect classifier h_c^{perf}, and thus needs (see the discussion for the **MON** axiom) to be given the highest possible score. By the same token, when $AN = 0$

the trivial rejector h_c^{rej} is the pervert classifier h_c^{perv}, and thus needs to be given the lowest possible score. Analogous arguments apply to the $AP = 0$ case.

AXIOM 8. ***Symmetry (SYM)****. For any test set* $\langle D, \mathcal{Y}_c \rangle$ *and for any classifier* h_c*,* $M(\mathcal{Y}_c, h_c) = M(\overline{\mathcal{Y}}_{\overline{c}}, \overline{h}_{\overline{c}})$ *holds.* □

SYM enforces the notion that the evaluation measure should be invariant with respect to switching the roles of the class c and its complement $\overline{c}$. This is desirable because it is not always the case that binary classification is naturally understood as the choice between a class and "its complement" (e.g., webpages about nuclear waste disposal vs. webpages not about it), where members of the first are naturally interpreted as "the positives". Sometimes the natural interpretation is a choice between two classes of equal standing (e.g., Shakespeare vs. Marlowe; Endorsements vs. Rebuttals; ProDemocrat vs. ProRepublican; FakeReviews vs. AuthenticReviews; Spam vs. Legitimate; etc.). In this case, it would be undesirable for the measure to return different results depending on which of the two is taken to be "the class" and which is taken to be "the complement of the class".

3.2 Discussion

SDE, **WDE** and **CON** deserve some comment, as they are mutually dependent. In a sense, all measures can be made to satisfy **SDE** and **WDE** by stipulating, for the cases in which they are (strongly or weakly) undefined, specific values that they should take up. For instance, the equation that defines F_1 (see Table 2) is such that F_1 is undefined when all of TP, FP, FN are 0, which means that F_1 would satisfy neither **SDE** nor **WDE**; in this case we may simply stipulate that, say, when $TP = FP = FN = 0$ then $F_1 = 1$, so that **SDE** and **WDE** are satisfied. The problems with this approach are that (a) when researchers propose or use an evaluation measure, they often omit to say what its output values are meant to be for the input values that make the function undefined, and (b) even when these output values are specified, they may generate points of discontinuity, i.e., make **CON** unsatisfied (see the discussion on F_1 and **CON** in Section 4). For the reasons above, in the next sections we will mostly concentrate on axioms other than **SDE** and **WDE**, since those other axioms do not have easy "fixes" as **SDE** and **WDE**.

One might think that the emphasis on axioms such as **SDE**, **WDE**, **CON** is excessive, since failure to satisfy them usually derives from the behaviour of the function in limiting cases, e.g., when $TP = FP = FN = 0$ and $TN = |D|$. We think that this emphasis is not excessive, since these limiting cases occur quite frequently in practice. For instance, in the well-known MLMCC Reuters-21578 collection[4], out of the 115 classes normally used for experimentation by researchers, no less than 25 are such that $AP = 0$. When results are macroaveraged (i.e., expressed as an unweighted average across the classes), Reuters-21578 results are determined for $\approx 21.7\%$ (since $25/115 \approx 21.7$) by classes such that $AP = 0$. In this, Reuters-21578 is not an exception, since large classification schemes usually exhibit a power-law behaviour, i.e., they typically consists of a few high-frequency classes and very many low- or very-low-frequency classes.

Another aspect that deserves mentioning is that not all axioms are equally desirable, since the motivations that lie behind these axioms are not all equally compelling. For instance, Axiom 8 (**SYM**) is desirable but probably not of fundamental importance, while Axiom 1 (**MON**) is of so fundamental importance as to invalidate, in our opinion, a measure that does not satisfy it. However, we will not attempt any classification of these axioms as "important vs. unimportant", since this is arguably a matter of degree.

[4]`http://bit.ly/1F8AFcO`

4. PROPERTIES OF THE F1 MEASURE

The F_1 measure is the most widely adopted evaluation metric for binary classification. In binary *text* classification F_1 has been the dominant measure ever since the recall-precision breakeven measure was deprecated in the late '90s [5]. The use of F_1 in text classification was first proposed in [16] (see also [15] for more on F_1 in text classification).

F_1 is based on the E_α measure, introduced by van Rijsbergen [27] and defined as

$$E_\alpha = 1 - \frac{1}{\alpha\frac{1}{\pi} + (1-\alpha)\frac{1}{\rho}} \qquad 0 \leq \alpha \leq 1 \tag{1}$$

where α is a parameter whose role is to specify the relative importance of precision and recall; a value $\alpha = 1/2$ attributes them equal importance. Note that E_α is a measure of error, not of accuracy, so lower values of E_α are better. F_1 is defined as

$$\begin{aligned} F_1 = 1 - E_{\frac{1}{2}} &= \frac{2\pi\rho}{\pi+\rho} = \frac{2TP}{2TP+FP+FN} \\ &= \frac{2TP}{AP+AN+TP-TN} \end{aligned} \tag{2}$$

where the last passage makes explicit the dependence of F_1 on the two variables (TP and TN) and two constants (AP and AN) of our problem.

We will now discuss how F_1 copes with respect to some of the axioms of Section 3.

PROPERTY 4.1. *F_1 does not satisfy Axiom 1 (**MON**).*

PROOF. Let us examine the case in which $TP = 0$ and $FN > 0$. In this case $F_1 = 0$ regardless of the values of FP; e.g., $TN = AN$ and $FP = 0$ (all the actual negatives have been classified correctly) and $TN = 0$ and $FP = AN$ (all the actual negatives have been misclassified) return the same result, i.e., $F_1 = 0$. This shows that F_1 fails to comply with Axiom 1. □

PROPERTY 4.2. *F_1 does not satisfy Axiom 2 (**CON**).*

PROOF. As from its definition, $F_1 = 1$ when $TP = FN = FP = 0$. However, when $TP = FN = 0$, it holds that

$$\lim_{FP \to 0} \frac{2TP}{2TP+FP+FN} = 0$$

which shows that F_1 is discontinuous at $TP = FN = FP = 0$, which proves our proposition. That $TP = FN = FP = 0$ is a problematic case for F_1 is also shown by the fact that

$$\begin{aligned} \frac{\partial F_1}{\partial TP} &= \frac{2(AP+AN-TN)}{(AP+AN+TP-TN)^2} \\ \frac{\partial F_1}{\partial TN} &= \frac{2(AP+AN+TP)}{(AP+AN+TP-TN)^2} \end{aligned} \tag{3}$$

[5]See Footnote 19 of [23] for a discussion of this point.

are both undefined when $(AP + AN + TP - TN) = 0$, i.e., when $TP = AP = 0$ and $TN = AN = |D|$, which proves our proposition again. □

This problem is reflected in the fact that what should F_1 be taken to return when $TP = FP = FN = 0$ and $TN = |D|$ is controversial. Some researchers (see e.g., [11]) maintain that in this case F_1 should evaluate to 1, since the classifier has classified all items correctly; incidentally, unless this is the case, F_1 does not satisfy **MON**. Other researchers have F_1 evaluate to 0 (e.g., [17]), likely on the grounds that, when $\rho = 0$, F_1 returns 0 for all other values of TN; note that this latter is a "continuity argument", applied to a situation in which (as we have seen) F_1 is not continuous. Yet other researchers (e.g., [15]) maintain that more than one value could be legitimate. To make matters worse, most other researchers do not actually specify, when using F_1, how they handle this case, which makes the results they report (especially those framed in terms of "macroaveraged F_1" in MLMCC) difficult to interpret and to compare with other results on the same datasets.

PROPERTY 4.3. *F_1 does not satisfy Axiom 6 (**CHA**).*

PROOF. It is easy to check that different ground truths generally give rise to different values of $E[F_1(\mathcal{Y}_c, h_c^{rand})]$. For instance, assume that $\langle D, \mathcal{Y}_c\rangle$ is such that $AN = 0$; if $|D| = 1$ then $E[F_1(\mathcal{Y}_c, h_c^{rand})] = 0.500$, while if $|D| = 100$ then $E[F_1(\mathcal{Y}_c, h_c^{rand})] \approx 0.612$. □

PROPERTY 4.4. *F_1 does not satisfy Axiom 7 (**IMB**).*

PROOF. Assume $AP > 0$ and $AN > 0$. For the trivial acceptor h_c^{acc} it holds that $TP = AP$, $FP = AN$, $FN = 0$, which means that $F_1(\mathcal{Y}_c, h_c^{acc}) = 2AP/(2AP + AN)$; this is not constant across all test sets, since it depends on the relative cardinalities of AP and AN[6]. □

The fact that F_1 does not satisfy **IMB** has undesirable consequences in terms of the interpretability of its results. For instance, is an F_1 value of 0.70 "good"? Most practitioners would answer "Yes", and this is indeed a good result if the relative frequency of class c is, say, 0.01 (in this case, $F_1(\mathcal{Y}_c,\ h_c^{acc}) = 2AP/(2AP + AN) \approx 0.01$), but cannot be considered a good result when the prevalence (i.e., relative frequency) of c is, say, 0.60, since in this case $F_1(\mathcal{Y}_c, h_c^{acc}) = 2AP/(2AP + AN) = 0.75$; i.e., in the latter case $F_1 = 0.70$ is well below the value obtained by a trivial classifier on the same data! Note that cases in which the prevalence of the class is 0.60 are not uncommon (as in all the cases mentioned at the end of Section 3.1), and that a perfectly balanced problem (i.e., when the relative frequency of c is 0.5) gives rise to $F_1(\mathcal{Y}_c, h_c^{acc}) \approx 0.666$.

The fact that F_1 does not satisfy **IMB** is extremely surprising, since F_1 is usually considered robust to imbalance, and is indeed the measure of choice for imbalanced binary classification. The reason of this apparent contradiction is that F_1 is considered robust to imbalance simply because, in the presence of imbalanced data, h_c^{acc} and h_c^{rej} return "low" values. However, (i) this occurs only when c is the minority class (see below about F_1 not satisfying Axiom 8), (ii) the values returned by h_c^{acc} and h_c^{rej} are not constant, and strongly depend on the prevalence of c, and (iii) these values increase steeply as the prevalence of c increases. In imposing **IMB** we are stating that, for a measure to be robust to imbalance, it is not enough that h_c^{acc} and h_c^{rej} return low values when the prevalence of c is low: the most important fact is that these values must *always be the same*, and independent of the prevalence of c.

[6] $F_1(\mathcal{Y}_c, h_c^{rej})$ is instead a constant independent of $\langle D, \mathcal{Y}_c\rangle$, since it is always 0 whenever $AP > 0$; in fact, when $TP = FP = 0$ and $FN = AP$ then $F_1 = \frac{0}{AP} = 0$.

PROPERTY 4.5. *F_1 does not satisfy Axiom 8 (**SYM**).*

PROOF. In switching from h_c to $\overline{h_{\overline{c}}}$ and from $\mathcal{Y}_c$ to $\overline{\mathcal{Y}_{\overline{c}}}$, TP and TN switch their roles, as do FP and FN. That F_1 does not satisfy **SYM** is thus shown by simply observing that

$$\frac{2TP}{2TP + FP + FN} \neq \frac{2TN}{2TN + FN + FP} \qquad \square$$

4.1 Other measures for classification

Like F_1, all of the measures listed in Table 2 fail to satisfy some fundamental axiom. Some examples are listed below.

PROPERTY 4.6. *ASP, DOR, $LAM\%$ do not satisfy Axiom 1 (**MON**).*

PROOF. Similarly to F_1, if $TP = 0$ and $FN > 0$ then ASP, DOR, $LAM\%$ take up values that are independent of how the actual negatives distribute across the false positives and the true negatives. While this suffices to prove our statement, note that for DOR and $LAM\%$ the same also holds when $TN = 0$; $LAM\%$ is also such that, when either FP or FN are 0, its value is the same irrespective of the value of the other among FP and FN. □

PROPERTY 4.7. *ASP, MCC and $LAM\%$ do not satisfy Axiom 2 (**CON**).*

PROOF. The partial derivatives of ASP with respect to variables TP and TN are

$$\begin{aligned}\frac{\partial ASP}{\partial TP} &= \frac{TP(2AP + 2AN - 2TN + TP)}{(AP + AN + TP - TN)^2}\\ \frac{\partial ASP}{\partial TN} &= \frac{TP^2}{(AP + AN + TP - TN)^2}\end{aligned} \tag{4}$$

These two derivatives are both undefined when $(AP+AN+TP-TN) = 0$, i.e., when $TP = AP = 0$ and $TN = AN = |D|$, which shows that ASP is not in C^1.

Concerning $LAM\%$, both derivatives $\partial LAM\%/\partial TP$ and $\partial LAM\%/\partial TN$ (not reported here since they are too complex) are undefined for both $TP = 0$ and $TN = 0$, which shows that $LAM\%$ is not in C^1

Concerning MCC, both $\partial MCC/\partial TP$ and $\partial MCC/\partial TN$ (also not reported here since they are too complex) are undefined for $TP + TN = |D|$, which shows that MCC is not in C^1. □

PROPERTY 4.8. *ASP and MCC do not satisfy Axiom 3 (**SDE**).*

PROOF. ASP is undefined for $TP = AP = 0$, since in this case it evaluates to $\frac{0}{0}$. MCC is undefined for either $AP = 0$ or $AN = 0$, since in this case it evaluates to $\frac{0}{0}$. □

PROPERTY 4.9. *DOR and $LAM\%$ do not satisfy Axiom 4 (**WDE**).*

PROOF. Assume we deal with a certain $\langle D, \mathcal{Y}_c\rangle$ such that $AP > 0$ and $AN > 0$. In this case (a) DOR is defined for any classifier h_c such that $FP > 0$ and $FN > 0$, but is not defined for all classifiers h'_c such that $FP = 0$ or $FN = 0$; and (b) $LAM\%$ is defined for most cases in which both $TP > 0$ and $TN > 0$ but is undefined for all cases in which either $TP = 0$ and $TN = 0$. □

PROPERTY 4.10. *Accuracy and ASP do not satisfy Axiom 7 (**IMB**).*

PROOF. For the trivial acceptor h_c^{acc}, since $AP = TP$ and $TN = 0$, then $Acc = \frac{TP+TN}{|D|} = \frac{AP}{|D|}$, and since it is also true that $PP = |D|$, then $ASP = \frac{TP^2}{AP\ PP} = \frac{AP}{|D|}$. So, in this case both accuracy and ASP coincide with the relative class frequency $\frac{AP}{|D|}$ of the class; therefore, in general they are different for different test sets. □

Note that, concerning DOR and $LAM\%$, we cannot even say whether they satisfy **IMB** or not, since for h_c^{acc} and h_c^{rej} they are not even defined.

5. PROPERTIES OF THE K MEASURE

In the previous sections we have seen that all of the measures listed in Table 2, including F_1, are unsatisfactory, since they all fail to satisfy one or more fundamental axioms among the ones we have argued for. As a measure of effectiveness for binary classification we then discuss K, which we define as

$$K = \begin{cases} \rho + \sigma - 1 & \text{if } AP > 0 \text{ and } AN > 0 \\ 2\sigma - 1 & \text{if } AP = 0 \\ 2\rho - 1 & \text{if } AN = 0 \end{cases} \tag{5}$$

where $\rho = TP/AP$ denotes recall and $\sigma = TN/AN$ denotes specificity. That is, when recall and specificity are both defined, K is a rescaled sum of recall and specificity; when one of them is not defined, K coincides with a rescaled version of the other. K is not entirely new, since it is a variant of

- *Youden's index* [29], or *informedness* [21], defined as $(\rho + \sigma - 1)$;
- *balanced accuracy* [4, 12, 24], defined as $(\rho + \sigma)/2$.

The main difference between K and these measures is that the proposers of the latter do not discuss exactly how to extend them to the cases in which either ρ or σ are undefined; how these extensions are accomplished impacts on the axioms that the measure does or does not satisfy.

Let us analyse the behaviour of K in the three cases listed in Equation (5). When $AP > 0$ and $AN > 0$ we have

$$K = \rho + \sigma - 1 = \frac{TP\ AN + TN\ AP}{AP\ AN} - 1 \tag{6}$$

When there are no positives ($AP = 0$) recall is undefined; in this case we let K default to specificity, since when there are no positives the system's ability to avoid false negatives is a non-problem, and the best the system can do is to correctly recognize all the negative examples as such, i.e., maximize specificity. Similarly, when there are no negatives ($AN = 0$) specificity is undefined, and we let K default to recall.

An evaluation measure for binary classification must reward the ability of the system to avoid false positives *and* the ability of the system to avoid false negatives. Similarly to F_1, K measures the ability of the system to avoid false negatives by means of recall; differently from F_1, K measures the ability of the system to avoid false positives by means of specificity (F_1 measures it by means of precision).

Let us now check how K behaves with respect to the axioms laid out in Section 3.1.

PROPERTY 5.1. *K satisfies Axiom 1 (**MON**).*

PROOF. Assume that h'_c differs from h_c only for the label attributed to a single item $d \in D$, wrong for h_c and correct for h'_c. If d is a false negative for h_c then it is a true positive for h'_c, which means that $\rho(h_c) < \rho(h'_c)$ and $\sigma(h_c) = \sigma(h'_c)$; if d is a false positive for h_c then it is a true negative for h'_c, which means that $\rho(h_c) = \rho(h'_c)$ and $\sigma(h_c) < \sigma(h'_c)$. In both cases it derives that $K(\mathcal{Y}_c, h_c) < K(\mathcal{Y}_c, h'_c)$. □

PROPERTY 5.2. *K satisfies Axiom 2 (**CON**).*

PROOF. If $AP = 0$ (resp., $AN = 0$), then $K = (2\sigma - 1)$ and $\partial K/\partial TN = 2/AN$ (resp., $K = (2\rho - 1)$ and $\partial K/\partial TP = 2/AP$), which is a constant. If $AP > 0$ and $AN > 0$, then $\partial K/\partial TP = 1/AP$ and $\partial K/\partial TN = 1/AN$, both also constants. This proves that K is in C^1. □

PROPERTY 5.3. *K satisfies Axiom 3 (**SDE**).*

PROOF. Trivial. □

PROPERTY 5.4. *K satisfies Axiom 4 (**WDE**).*

PROOF. Follows from **SDE**, since **SDE** strictly implies **WDE**. □

PROPERTY 5.5. *K satisfies Axiom 5 (**FIX**).*

PROOF. If $AP = 0$ (resp., $AN = 0$), then $K = (2\sigma - 1)$ (resp., $K = (2\rho - 1)$) ranges on the [-1,+1] interval. If $AP > 0$ and $AN > 0$, then $K = \rho + \sigma - 1$ ranges on [-1,+1] since both ρ and σ range on [0,1] and are independent. In particular, the perfect classifier has a value of $K = 1$, since $\rho = 1$ and $\sigma = 1$, and the pervert classifier has a value of $K = -1$, since $\rho = 0$ and $\sigma = 0$. So, K always ranges on [-1,+1] irrespectively of the test set $\langle D, \mathcal{Y}_c \rangle$. □

PROPERTY 5.6. *K satisfies Axiom 6 (**CHA**).*

PROOF. For every test set $\langle D, \mathcal{Y}_c \rangle$, for every classifier h_c there is a unique classifier h'_c such that $h_c(d) = -h'_c(d)$; this latter is such that $K(\mathcal{Y}_c, h_c) = -K(\mathcal{Y}_c, h'_c)$, so the mean of the K scores of h_c and h'_c is 0. Since h_c and h'_c are equiprobable, it derives that $E[K(\mathcal{Y}_c, h_c^{rand})] = 0$ for each test set $\langle D, \mathcal{Y}_c \rangle$. □

PROPERTY 5.7. *K satisfies Axiom 7 (**IMB**).*

PROOF. If $AP > 0$ and $AN > 0$, then $K(\mathcal{Y}_c, h_c^{acc}) = 0$, since $\rho(h_c^{acc}) = 1$ and $\sigma(h_c^{acc}) = 0$, while $K(\mathcal{Y}_c, h_c^{rej}) = 0$, since $\rho(h_c^{rej}) = 0$ and $\sigma(h_c^{rej}) = 1$. □

PROPERTY 5.8. *K satisfies Axiom 8 (**SYM**).*

PROOF. This is shown by noting that $\sigma(h_c) = \rho(\overline{h}_{\overline{c}})$ and $\rho(h_c) = \sigma(\overline{h}_{\overline{c}})$, and by noting that K is symmetric with respect to ρ and σ. □

5.1 Discussion

We have seen that, while F_1 fails to comply with a number of axioms (**MON**, **CON**, **CHA**, **IMB**, **SYM**), K satisfies all the eight axioms we have argued for in Section 3. While this shows the superiority of K over F_1, there are additional reasons why the former should be preferred to the latter:

1. K is based on two *independent* quantities (recall and specificity), while F_1 is based on two dependent quantities, precision and recall (one cannot increase recall

without also increasing precision[7]), which is odd. That recall and specificity are independent can be seen by the fact that they are computed on two non-overlapping halves of the contingency table (TP and FN for recall, TN and FP for specificity), while recall and precision are computed on two *overlapping* halves (TP and FN for recall, TP and FP for precision).

2. K takes all the elements of the contingency table into account, while this is not true for F_1, which seems especially unsuitable when c and $\overline{c}$ are two classes of equal standing (e.g., ProDemocrat vs. ProRepublican). For instance, given two contingency tables $t_1 = \langle TP, FP, FN, TN\rangle$ and $t_2 = \langle TP, FP, FN, 1000000 * TN\rangle$, F_1 is the same for both t_1 and t_2 (which is odd), while this is not true for K.

3. It is linear. It is thus easy (much easier than, say, F_1 or $LAM\%$) to use as a loss function that gets explicitly minimized within supervised learning algorithms.

4. It is extremely simple. This means it can be easily understood even by people with little mathematical background (e.g., company managers), for whom even the very notion of "harmonic mean" present in the definition of F_1 is esoteric.

6. EXTENDING K TO SLC, CSC, AND OC

We now turn our attention to classification problems other than binary classification. Classification problems may be ordered according to a "specialization hierarchy", where

- *Binary classification* (BC) is a special case of *single-label classification* (SLC). SLC is defined as the task of assigning to each item exactly one class from a set $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$, where $|\mathcal{C}| > 1$. BC corresponds to the $|\mathcal{C}| = 2$ case[8], while *single-label multi-class classification* (SLMCC) corresponds to the $|\mathcal{C}| > 2$ case.

- Both SLC and *ordinal classification* (OC) are special cases of *cost-sensitive classification* (CSC), defined as the task of assigning to each item exactly one class from a set of classes $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$, where $|\mathcal{C}| > 1$ and where a set of pairwise *distances* (or *costs*) $\Delta(c_i, c_j) \geq 0$ between classes is defined such that

  - $\Delta(c_i, c_i) = 0$ for all $c_i \in \mathcal{C}$
  - $\Delta(c_i, c_j)$ quantifies the cost of misclassifying into c_i an item which actually belongs to c_j.

  The set of $\Delta(c_i, c_j)$ values is usually referred to as the *cost matrix.* Accordingly,

  - SLC is the case in which $\Delta(c_i, c_j) = 1$ for all $c_i, c_j \in \mathcal{C}, i \neq j$;
  - OC is the case in which $|\mathcal{C}| > 2$ and, for all $c_i, c_j \in \mathcal{C}$ such that $i < j$, it holds that

$$\begin{aligned} \Delta(c_i, c_j) &= \Delta(c_j, c_i) \\ \Delta(c_i, c_j) &= \sum_{k=i}^{j-1} \Delta(c_k, c_{k+1}) \end{aligned} \tag{7}$$

A problem with current evaluation measures for classification is that they do not reflect the specialization hierarchy of classification problems. For instance, while F_1 is a standard measure used for evaluating binary classification, there is no known equivalent of F_1 for CSC or OC. In the following we define such equivalent for K, i.e., extend K to cover the general CSC case (hence the SLMCC and OC cases too); this means that K can be used as a unifying evaluation measure for all types of classification.

6.1 The SLC case

We start by addressing SLC. In order to discuss this case, let's fix some notation. By TP_j, FP_j, FN_j, and TN_j we will indicate the numbers of true positives, false positives, false negatives, and true negatives, *for class* c_j; for instance, FN_j will indicate the number of items that belong to class c_j and were instead predicted to belong to some class different from c_j. AP_j and AN_j are defined accordingly.

Let us define the indicator variable

$$\xi_j = \begin{cases} 1 & \text{if } AP_j > 0 \\ 0 & \text{if } AP_j = 0 \end{cases} \tag{8}$$

and let us define *recall for* c_j (indicated as ρ_j) as

$$\rho_j = \begin{cases} \dfrac{TP_j}{AP_j} & \text{if } AP_j > 0 \\ undefined & \text{if } AP_j = 0 \end{cases} \tag{9}$$

Note that, in the binary case, $\sigma(h_c)$ is equivalent to $\rho(\overline{h_{\overline{c}}})$, hence K may be viewed as (a rescaled version of) the sum of the recall values for the two binary classes c and $\overline{c}$. This suggests a natural extension of K to the SLC case, as

$$K = \frac{|\mathcal{C}|}{|\mathcal{C}|-1} \frac{\sum_{c_j \in \mathcal{C}, \xi_j = 1} \rho_j}{\sum_{c_j \in \mathcal{C}} \xi_j} - \frac{1}{|\mathcal{C}|-1} \tag{10}$$

which is a rescaled variant of *macroaveraged recall.* It is easy to observe that, when $|\mathcal{C}| = 2$, Equation (10) defaults to Equation (5). It is also easy to check that all the axioms discussed in Section 3, that we proved to hold for the "binary" version of K, also hold for this "multiclass" version.

6.2 The CSC case and the OC case

We may extend K to the general cost-sensitive classification case (and hence to the ordinal classification case). CSC (see e.g., [9, 10]) is important in many real-life applications (e.g., spam filtering, medical diagnosis) in which some classification errors have more serious consequences than others. OC (also known as *ordinal regression* – see e.g., [6, 26]) is also important due to its key role in the social sciences, where ordinal (i.e., discrete) scales are often used to elicit human judgments and evaluations from respondents or interviewees.

We extend K to deal with cost-sensitive classification by defining a notion of recall that is sensitive to the error $E(d_i)$ made in misclassifying an item d_i into a class $h_c(d_i)$ that has a certain distance $\Delta(h_c(d_i), \mathcal{Y}_c(d_i))$ from its true class $\mathcal{Y}_c(d_i)$. $E(d_i)$ may be one of the metrics popular in ordinal

[7] Assume a classifier h'_c identical to h_c except it has one less false negative and one more true positive; in moving from h_c to h'_c, recall has increased, but precision has also increased, since TP has increased and FP is unmodified.

[8] *Multi-label multi-class classification* (MLMCC) is defined as the task of assigning to each item zero, one, or several classes from a set of classes $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$, where $|\mathcal{C}| > 1$. As such, MLMCC is equivalent to BC (at least from the standpoint of evaluation), since it corresponds to performing BC independently for each of the classes in $\mathcal{C}$. We will thus not consider it a separate task.

classification, such as absolute error

$$AE = \Delta(h_c(d_i), \mathcal{Y}_c(d_i))$$

or squared error

$$SE = \Delta(h_c(d_i), \mathcal{Y}_c(d_i))^2$$

Let us define recall on class c_j as

$$\rho_j = \begin{cases} \dfrac{\sum_{i=1}^{AP_j}(1 - \dfrac{E(d_i)}{\max(E_j)})}{AP_j} & \text{if } AP_j > 0 \\ undefined & \text{if } AP_j = 0 \end{cases} \quad (11)$$

Here, $\max(E_j)$ is the maximum possible error that could be made in misclassifying an item whose true class is c_j (i.e., the error that we make in picking the class most distant from true class c_j). It can be easily checked that ρ_j is 1 if and only if all items belonging to c_j are correctly classified into c_j, and is 0 if and only if all items belonging to c_j are misclassified with the maximum possible error, i.e., into the class most distant from c_j. As such, ρ_j is a natural extension of the notion of recall as we know it from binary classification.

EXAMPLE 6.1. *Assume that absolute error AE is our measure of error E, that* $C = \{c_1, ..., c_5\}$*, that* $\Delta(c_i, c_{i+1}) = 1$ *for all* $i \in \{1, 2, 3, 4\}$*, that items* d_1*,* d_2*,* d_3 *all have true class* c_3*, that* d_1 *is correctly classified into* c_3*, that* d_2 *is misclassified into* c_2*, and that* d_3 *is misclassified into* c_1*. Assume that item* d_4 *has true class* c_4 *and is misclassified into* c_3*.*

The contribution of d_1 *to* ρ_3 *is* $(1 - 0) = 1$*, while the contribution of* d_2 *is* $(1 - \frac{1}{2}) = \frac{1}{2}$ *and the contribution of* d_3 *is* $(1 - \frac{2}{2}) = 0$*; the contribution of* d_4 *to* ρ_4 *is* $(1 - \frac{1}{3}) = \frac{2}{3}$*.*

While both d_2 *and* d_4 *are misclassified into a class with distance 1 from their true class, the error made for* d_2 *is considered more severe than that made for* d_4*, since error is evaluated relative to the maximum possible error, which is different for different classes* c_j*.* □

For CSC we stick to the definition of K, unchanged, as given in Equation (10); the difference with the SLC case is thus in the notion of recall adopted (Equation (11) instead of Equation (9)), and not in the way of summing the class-specific values of recall, which remains the same as in standard SLC.

It is immediate to check that if distances have all the same magnitude, i.e., $\Delta(c_i, c_j) = 1$ for all $c_i, c_j \in \mathcal{C}, i \neq j$, as in standard SLC, Equation (11) defaults to Equation (9). It is also easy to check that all the axioms that we have shown to hold for the binary and SLC versions of K, also hold for the CSC version (and, as a consequence, for the OC version).

7. RELATED WORK

This axiomatic approach to evaluating evaluation measures is not new in IR. For instance, [2] studies measures for evaluating clustering systems axiomatically, while [19] does the same for measures for evaluating ad hoc search. Sokolova and Lapalme [25] discuss properties of classification measures, but focus on properties of invariance across test sets characterised by different sets D, which is hardly of interest to the present context.

More recently, in discussing where the "Frontiers, Challenges, and Opportunities for Information Retrieval" lie, the SWIRL 2012 participants [1, p. 20] called for the development of *axiometrics* for IR (see also [5]), i.e., axiomatically defined evaluation metrics. It is exactly axiometrics for classification that we are looking at here. The effort closest in spirit to the present one is [3], which proposes axiomatic studies of evaluation measures for filtering systems (especially focusing on cost-sensitive measures); since filtering is an instance of classification, [3] is relevant to the present work. In [3] the authors claim that the main difference between metrics is how h_c^{acc}, h_c^{rej}, h_c^{rand} are evaluated, and only identify two axioms that they argue should be satisfied by any evaluation measure; these axioms are **MON** plus another weaker axiom, strictly entailed by **MON**, which says that only the perfect classifier can obtain the highest M score. One further difference between [3] and the present work is that [3] has a *descriptive* intent, i.e., describes a number of axioms but does not necessarily argue that a measure should satisfy them; our work has a *normative* character instead, i.e., we describe a number of axioms *and* argue that a worthwhile measure should satisfy them.

We should recall that an early mention of the axiomatic approach to evaluating binary retrieval is to be found in van Rijsbergen's work [27, 28], where the author discusses a number of formal properties (that collectively characterize "additive conjoint structures") that, as he argues, combinations of precision and recall should satisfy. The author goes on to propose one such combination (the E_α measure of Equation 1) but does not prove that it indeed satisfies the said formal properties. Binary retrieval and binary classification are strongly related, so van Rijsbergen's work is indeed relevant to our quest. However, our approach is more general than his, since he focuses on properties that a combination of two simple measures (precision and recall, in his case) should satisfy, while the properties we study view the evaluation measure as a direct function of the contingency table, without postulating (actually: without lending importance to) the presence of intermediate simple measures.

8. CONCLUSIONS

We have proposed K (a variant of "Youden's index" and "balanced accuracy") as an evaluation measure for binary classification. K has a number of interesting properties. The perfect classifier obtains $K = 1$, the pervert classifier obtains $K = -1$, the trivial acceptor and the trivial rejector both obtain $K = 0$, and the expected value of the random classifier is always $K = 0$; all of these hold irrespectively of class prevalence, which makes classification results expressed in terms of K easily interpretable. K is defined on all the cells of the contingency table, which makes it suitable for addressing both balanced and imbalanced test sets; in particular, this avoids the problem of defining what counts as a "balanced" test set. One advantage of K is that it smoothly extends to multi-label multi-class classification, single-label multi-class classification, cost-sensitive classification, and ordinal classification. K has the additional virtues of simplicity, which makes it easily interpretable by non-initiates, and linearity, which makes it easy to directly optimize by supervised learning algorithms.

We have obtained K via an "axiomatic" study of the properties that a measure for classification should have. This study has also shown that F_1, the currently standard evaluation measure for binary classification, is flawed, since it does not satisfy several properties that should intuitively hold for any satisfactory measure; of particular importance is the fact that F_1 is not monotonic and is not continuously differentiable. This hints at the power of the axiomatic approach, which we argue should be used more and more for scrutinizing the accepted wisdom in effectiveness evaluation.

Acknowledgments

I am grateful to Andrea Esuli, Tiziano Fagni, Donna Harman, David Lewis, Stefano Mizzaro, Steve Robertson, and Keith van Rijsbergen for interesting discussions about the subject of this paper. I am also indebted to one of the reviewers for spotting problems in previous formulations of Axioms 2 and 8.

9. REFERENCES

[1] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval. Report from SWIRL 2012. *SIGIR Forum*, 46(1):2–32, 2012.

[2] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[3] E. Amigó, J. Gonzalo, and F. Verdejo. A comparison of evaluation metrics for document filtering. In *Proceedings of the 2nd International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*, pages 38–49, Amsterdam, NL, 2011.

[4] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010)*, pages 3121–3124, Istanbul, TR, 2010.

[5] L. Busin and S. Mizzaro. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 4th International Conference on the Theory of Information Retrieval (ICTIR 2013)*, Copenhagen, DK, 2013.

[6] W. Chu and S. S. Keerthi. Support vector ordinal regression. *Neural Computation*, 19(3):145–152, 2007.

[7] G. V. Cormack and A. Kołcz. Spam filter evaluation with imprecise test set. In *Proceedings of the 32nd ACM Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 604–611, Boston, US, 2009.

[8] G. V. Cormack and T. Lynam. TREC 2005 Spam Track overview. In *Proceedings of the 14th Text Retrieval Conference (TREC 2005)*, Gaithersburg, US, 2005.

[9] P. M. Domingos. MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999)*, pages 155–164, San Diego, US, 1999.

[10] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 973–978, Seattle, US, 2001.

[11] A. Esuli and F. Sebastiani. Active learning strategies for multi-label text classification. In *Proceedings of the 31st European Conference on Information Retrieval (ECIR 2009)*, pages 102–113, Toulouse, FR, 2009.

[12] V. García, R. A. Mollineda, and J. S. Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. In *Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2009)*, pages 441–448, Póvoa de Varzim, PT, 2009.

[13] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. Bossuyt. The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135, 2003.

[14] D. Hull. The TREC-6 filtering track: Description and analysis. In *Proceedings of the 6th Text Retrieval Conference (TREC 1997)*, pages 33–56, 1997.

[15] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1995)*, pages 246–254, Seattle, US, 1995.

[16] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 3–12, Dublin, IE, 1994.

[17] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[18] B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.

[19] A. Moffat. Seven numeric properties of effectiveness metrics. In *Proceedings of the 9th Conference of the Asia Information Retrieval Societies (AIRS 2013)*, pages 1–12, Singapore, SN, 2013.

[20] D. W. Oard and W. Webber. Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, 7(2/3):99–237, 2013.

[21] D. M. Powers. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness, and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[22] S. E. Robertson. The parametric description of retrieval tests. Part I: The basic parameters. *Journal of Documentation*, 25(1):1–27, 1969.

[23] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[24] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence (AJCAI 2006)*, pages 1015–1021, Hobart, AU, 2006.

[25] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

[26] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li. Kernel discriminant learning for ordinal regression. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):906–910, 2010.

[27] C. J. van Rijsbergen. Foundations of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

[28] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, UK, second edition, 1979.

[29] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3:32–35, 1950.